Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W477–W481***
doi:10.1093/nar/gkl215

CAPweb: a bioinformatics CGH array Analysis Platform

Stéphane Liva[1,*], Philippe Hupé[1,2], Pierre Neuvial[1], Isabel Brito[1], Eric Viara[1], Philippe La Rosa[1] and Emmanuel Barillot[1]

[1]Institut Curie, Service Bioinformatique and [2]Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75248 Paris Cedex 05, France

Received February 14, 2006; Revised and Accepted March 24, 2006

ABSTRACT

Assessing variations in DNA copy number is crucial for understanding constitutional or somatic diseases, particularly cancers. The recently developed array-CGH (comparative genomic hybridization) technology allows this to be investigated at the genomic level. We report the availability of a web tool for analysing array-CGH data. CAPweb (CGH array Analysis Platform on the Web) is intended as a user-friendly tool enabling biologists to completely analyse CGH arrays from the raw data to the visualization and biological interpretation. The user typically performs the following bioinformatics steps of a CGH array project within CAPweb: the secure upload of the results of CGH array image analysis and of the array annotation (genomic position of the probes); first level analysis of each array, including automatic normalization of the data (for correcting experimental biases), breakpoint detection and status assignment (gain, loss or normal); validation or deletion of the analysis based on a summary report and quality criteria; visualization and biological analysis of the genomic profiles and results through a user-friendly interface. CAPweb is accessible at http://bioinfo.curie.fr/CAPweb.

INTRODUCTION

In recent years, array-CGH (comparative genomic hybridization) has become the technology of choice for large scale investigations of DNA copy number changes between two genomes. Today, CGH arrays allow the ratio of DNA copy number between a test and a reference sample to be simultaneously assessed in 2000 to 30 000 positions in the genome, giving a resolution of between 1.5 Mb to 100 kb (1,2). Its main applications are the study of diseases in which the DNA copy number varies in certain locations of the genomes, due to either constitutional mutations (hereditary or *de novo*), such as human genetic diseases (3) or somatic changes, such as in cancers (4). The identification of regions of altered DNA gives valuable information about the genes involved in the disease, and many projects have been launched worldwide to determine the genome structure of tumour cells (4). Array-CGH is also an important source of information for studying genome evolution, for example in bacteria (5) or mammals (6). We have developed a Web tool, called CAPweb (CAP: CGH array Analysis Platform), for bioinformatics analysis of CGH arrays. This tool combines the following tasks: (i) data management, (ii) array normalization, (iii) automatic breakpoint detection and assessment of gain and loss regions, (iv) quality control and (v) a graphical user interface for browsing and analysing the genomic profiles.

Several tools have recently been developed for analysing CGH array data, such as CGH-Explorer (7), ArrayCyGHt (8), CGHPRO (9), WebArray (10) or ArrayCGHbase (11), although the only web-accessible servers are ArrayCyGHt, WebArray and CAPweb. Among these three, only CAPweb allows project management and the upload of raw data files without pre-processing. It also offers unique features for the analysis and visualization of array-CGH data. CAPweb accepts raw data from the main microarray image analysis software. As far as we are aware, CAPweb is the only platform dedicated to biologists that allows the complete analysis of raw CGH arrays from the raw data to visualization and biological interpretation.

DESCRIPTION

The CAPweb server allows the user to store, analyse and manage his or her data. We will now describe its operation (Figure 1). A tutorial is accessible at http://bioinfo.curie.fr/tutorial/CAPweb/capweb_tutorial.html.

*To whom correspondence should be addressed. Tel: +33 0 1 4234 65 31; Fax: +33 0 1 42 34 65 28; Email: capweb@curie.fr

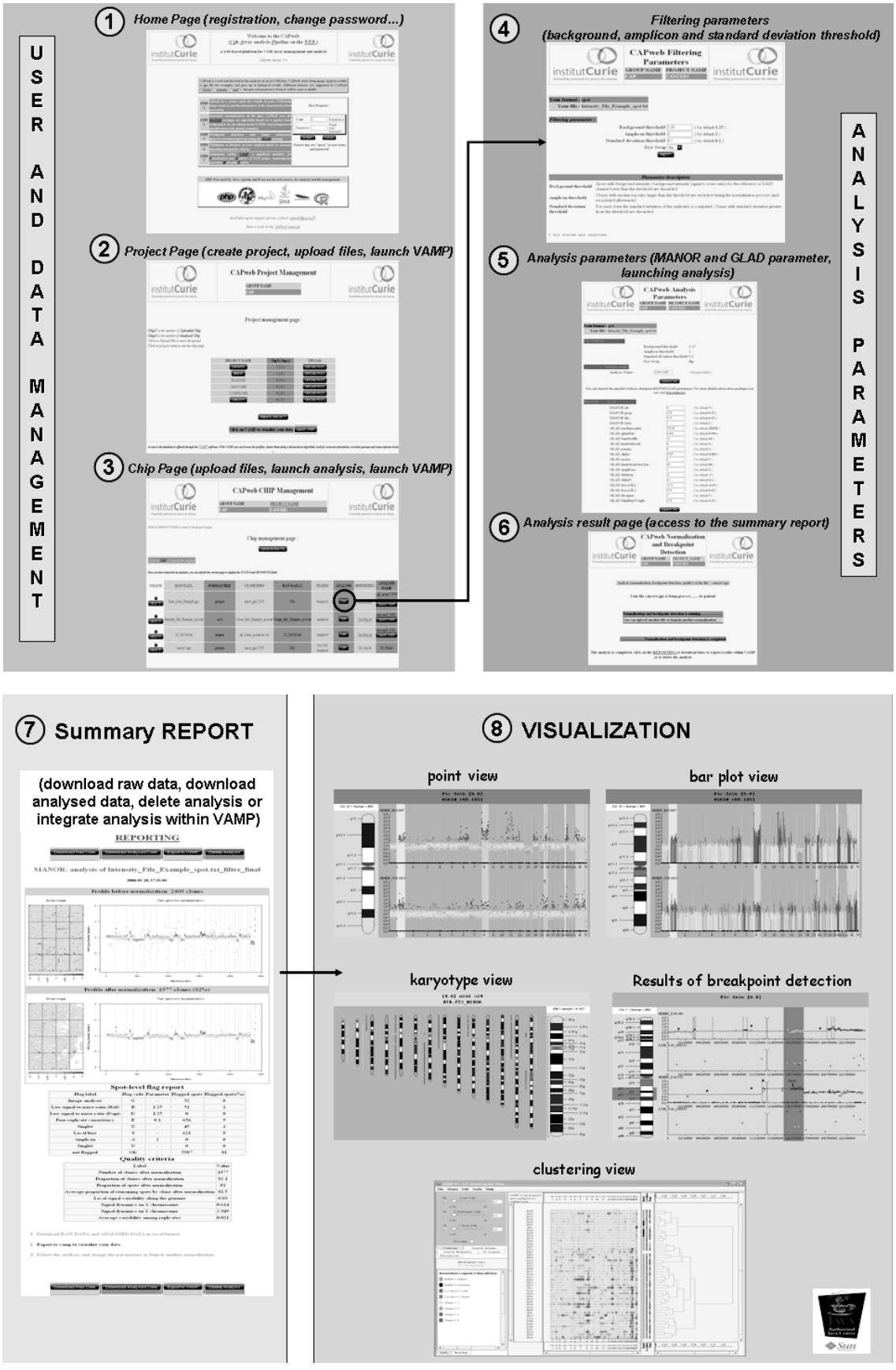

Figure 1. Different views of CAPweb Interface showing how the CGH array analysis proceeds, see text for details.

User registration, data upload and management

The first step of the analysis is user registration [Figure 1(1)], which ensures the confidentiality of the submitted data. The user is sent a login/password by email and can then create one or more projects to upload data files [Figure 1(2)]. Several input formats from microarray image analysis software are currently supported: Genepix (http://www.moleculardevices.com/pages/instruments/gn_genepix4000.html), Imagene (http://www.biodiscovery.com/index/imagene), Spot (12) and MAIA (13). CAPweb requires only two types of file: (i) a raw intensity file (one file for Genepix and MAIA, two files for Imagene and Spot) and (ii) a genomic position file mapping each spot to a name and its position on the genome under CSV (semi colon separator) format.

For each project, the 'Array Management' page [Figure 1(3)] lists all the arrays, their analysis status and the summary report file, and allows new analyses to be launched.

The array files are permanently stored on the server: the user can only browse the arrays of his or her projects, and only the user is allowed to delete them.

CGH array analysis

From the 'Array management' page, the user can launch the array analyses. The analyses are run in the background, allowing the user to use CAPweb for other analyses.

Data Normalization (MANOR). As in all microarray analyses, CGH array data must be normalized to correct for experimental artefacts while preserving the true biological signal. For this goal, CAPweb uses the Bioconductor package MANOR, which includes spot and clone filtering steps that discards spots having too low a signal-to-noise ratio or clones with a poor replicate consistency, and, most importantly, it includes a spatial normalization step. This step aims to correct for spatial effects on the arrays. We identified these as the predominant experimental artefact in the array-CGH data we have studied. The corresponding algorithm is based on a spatial trend estimation and a signal segmentation method with a spatial constraint, as described in P. Neuvial *et al*. (manuscript submitted).

Breakpoint detection and assessment of gain and loss region (GLAD). This step aims to identify chromosomal regions having an identical DNA copy number, which are delimited by breakpoints. CAPweb uses the Bioconductor package GLAD, which implements an algorithm described in (14). This method first uses the spatial structure of array-CGH data to adaptively calculate a smoothed signal value for each clone. These smoothed signal values are then used to detect breakpoints and outliers, and then genomic regions having the same underlying copy number are clustered together.

Quality control. Various statistical criteria can help the user assess the quality of the array. These include intra-replicate variability, genomic neighbour variability, the percentage of spots filtered out after image analysis and the amplitude of signal gap between regions having a different DNA copy number. These quality criteria are reported in an HTML summary report file, which also displays key features of the normalization process: array image and genomic profile before and after normalization, and a summary of the normalization. This file [Figure 1(7)] allows the user to compare the quality of the data before and after analysis. Based on this information, the user may choose to keep or discard the analysis.

This data analysis step can be run without an extensive knowledge of the underlying statistical algorithms by using default parameters. Default parameters have been calibrated by comparing quality criteria for various parameter value in two datasets: one from UCSF (218 arrays, Spot format, as a collaboration with Dan Pinkel), and one from Institut Curie/INSERM U509 (181 arrays, Genepix format). This part is described in detail elsewhere (P. Neuvial *et al*. manuscript submitted). However, CAPweb allows the user to choose the value of several parameters for filtering, spatial normalization and breakpoint detection. The summary report also helps in comparing the results of analyses carried out with different parameter values [Figure 1 (4–6)].

Visualization (VAMP) and biological analysis

Once the first level of array analysis has finished, the user can visualize and further analyse the data through a graphical user interface: VAMP—visualization and analysis of array-CGH, transcriptome and other molecular profiles (P. La Rosa *et al*. manuscript submitted) [Figure 1 (8)]. Several visualization types are proposed, such as the classical CGH karyotype view or the genome-wide multi-tumour comparison view. These allow the user to easily compare different arrays. Additional information concerning each clone or DNA region can be interactively retrieved from different public databases through external links. Other functions for analysing CGH data are provided within the interface, such as looking for minimal or recurrent regions of alterations (15), clustering, etc.

VAMP allows the user to display genomic profiles at various resolutions [from the whole genome to small regions (clone level)]. All the analyses results (breakpoint detection, assignment of gain/lost region, quality criteria, etc.) can also be displayed within VAMP. VAMP has many other functions for navigation, querying and analysis that we have not explained here; we refer the reader to the documentation and demo for further details (http://bioinfo.curie.fr/vamp/doc).

Note that the user can analyse at least 200 arrays with 1GB of memory.

IMPLEMENTATION

The CAPweb server is based on freely available components (Figure 2). The database for user management and array management was built on mySQL. PHP scripts ensure registration and project management. Perl scripts control the launching of statistical analyses written in R. A Java applet and XML files are used for the visualization. CAPweb integrates the MANOR and GLAD R packages and the VAMP software, all of which were developed at the Institut Curie.

The security in CAPweb is based on mysql authentication and cookie session. Uploaded data are considered strictly confidential. The CAPweb server is also available upon request for local installation on Unix/Linux/MacOS X operating systems.

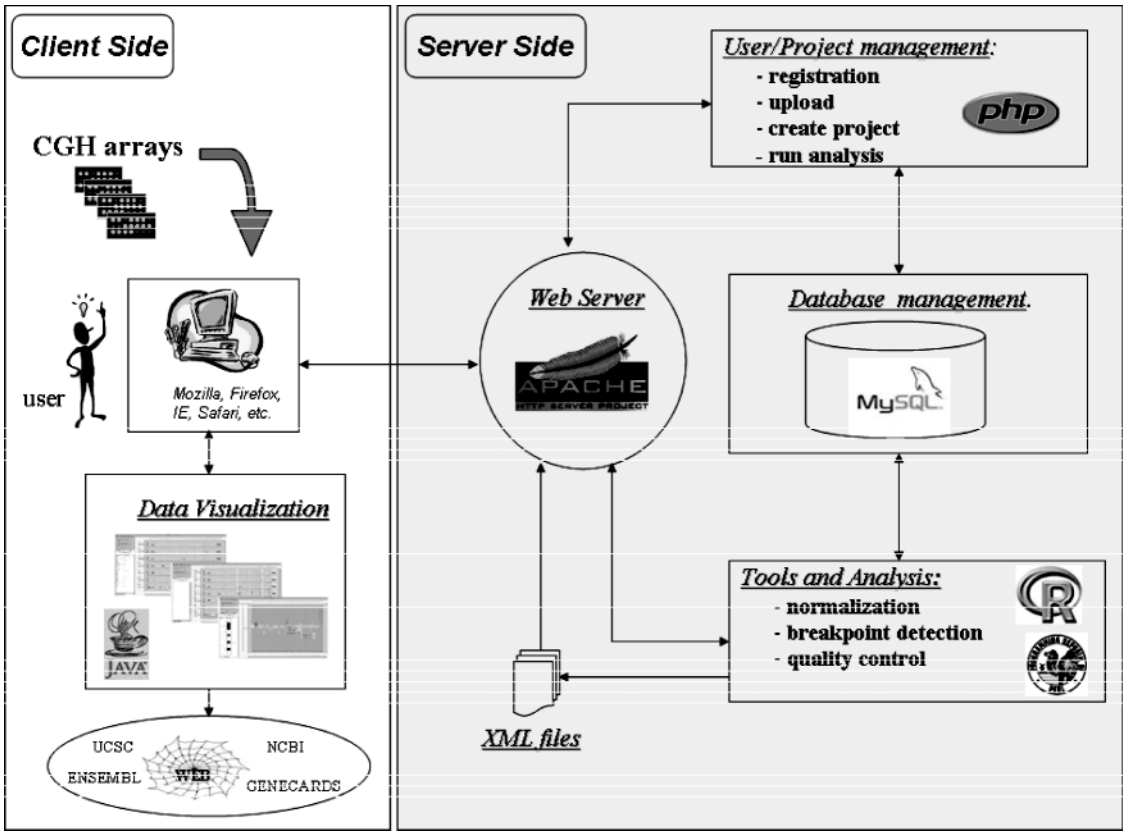

Figure 2. CAPweb software architecture, see text for details.

CONCLUSION

Array-CGH is a popular technology that is now used in many projects ranging from the characterization of tumours to the study of genome evolution. As with any large scale technology, its exploitation relies heavily on the availability of bioinformatics tools for managing and analysing the data. Many bioinformatics algorithms and interfaces have been developed but biologists have lacked a web-based platform for integrating these tools in a user-friendly manner. CAPweb offers this service and combines array normalization, quality control, breakpoint detection and the biological interpretation of the results. It also helps with data management. Currently, the public CAPweb server at the Institut Curie contains 800 arrays.

In this paper we have presented CAPweb 1.0 version. A new version is currently being developed, which will allow the user to analyse high density oligonucleotide arrays, such as Affymetrix GeneChip® Arrays or Nimblegen™ Arrays, to integrate any clinical information, and to add gene expression profiles so that copy number profiles can be compared and correlated to them.

ACKNOWLEDGEMENTS

This work was supported by the Institut Curie, the Centre National de la Recherche Scientifique, the Cancéropole Ile-de-France, the Région Ile-de-France and the association 'Courir pour la vie, Courir pour Curie'. The authors thank all our colleagues who have tested CAPweb and suggested improvements: G. Pierron, C. Brennetot, A. Idbaih, E. Manié (Institut Curie) and S. Law (UCSF). Funding to pay the Open Access publication charges for this article was provided by Institut Curie.

Conflict of interest statement. None declared.

REFERENCES

1. Snijders,A.M., Nowak,N., Segraves,R., Blackwood,S., Brown,N., Conroy,J., Hamilton,G., Hindle,A.K., Huey,B., Kimura,K. *et al.* (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet*., **29**, 263–264.
2. Ishkanian,A.S., Malloff,C.A., Watson,S.K., DeLeeuw,R.J., Chi,B., Coe,B.P., Snijders,A., Albertson,D.G., Pinkel,D., Marra,M.A. *et al.* (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genet*., **36**, 299–303.
3. Lockwood,W.W., Chari,R., Chi,B. and Lam,W.L. (2006) Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur. J. Hum. Genet*., **14**, 139–148.
4. Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nature Genet*., **37**, 11–17.
5. Fukiya,S., Mizoguchi,H., Tobe,T. and Mori,H. (2004) Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* Strains revealed by comparative hybridization microarray. *J. Bacteriol*., **186**, 3911–3921.
6. Wilson,G.M., Flibotte,S., Missirlis,P.I., Marra,M.A., Jones,S., Thornton,K., Clark,A.G. and Holt,R.A. (2006) Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. *Genome Res*, **16**, 173–181.
7. Lingjaerde,O.C., Baumbush,L.O., Liestol,K., Glad,I.K. and Borresen-Dale,A.L. (2005) CGH-explorer, a program for analysis of array-CGH data. *Bioinformatics*, **6**, 821–822.

8. Kim,S.Y., Nam,S.W., Lee,S.H., Park,W.S., Yoo,N.J., Lee,J.Y. and Chung,Y.J. (2005) ArrayCyGHt, a web application for analysis and visualization of array-CGH data. *Bioinformatics*, **21**, 2554–2555.
9. Chen,W., Erdogan,F., Ropers,H., Lenzner,S. and Ullmann,R. (2005) CGHPRO, a comprehensive data analysis tool for array CGH. *BMC Bioinformatics*, **6**, 85.
10. Xia,X., McClelland,M. and Wang,Y. (2005) WebArray, an online platform for microarray data analysis. *BMC Bionformatics*, **6**, 306.
11. Menten,B., Pattyn,F., De Preter,K., Robbrecht,P., Michels,E., Buysse,K., Mortier,G., De Paepe,A., van Vooren,S., Vermeesh,J. *et al*. (2005) ArrayCGHbase: an analysis platform for comparative genomic hybridization microarrays. *BMC Bioinformatics*, **6**, 124.
12. Jain,A.N., Tokuyasu,T.A., Snidjers,A.M., Segraves,R., Albertson,D.G. and Pinkel,D. (2002) Fully automatic quantification of microarray image data. *Genome Res*., **12**, 325–332.
13. Novikov,E. and Barillot,E. (2005) A robust algorithm for ratio estimation in two-color microarray experiments. *J. Bioinform. Comput. Biol*., **6**, 1411–1428.
14. Hupé,P., Stransky,N., Thiery,J.P., Radvanyi,F. and Barillot,E. (2004) Analysis of array CGH data: from signal ratio to gain and loss DNA regions. *Bioinformatics*, **20**, 3413–3422.
15. Rouveirol,C., Stransky,N., Hupé,P., La Rosa,P., Viara,E., Barillot,E. and Radvanyi,F. (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849–856.